Klaas Willems, Ghent University

Is frequency an explanatory causal concept in linguistics?

In the last two decades, the use of quantitative methods in synchronic and diachronic linguistics has been booming. Surely, this development is to be welcomed, if only because it has enriched the ways of thinking about various important problems of linguistic inquiry. Many linguists now agree that data on frequencies of occurrence serve more than just illustrative purposes. Anyone who has taken the trouble to study large amounts of language data in view of the subtleties of formal and semantic variation, both synchronically and diachronically, will readily admit that corpus-based research – and quantitative variationist research at large – reveals more than any intuition-based or introspection-based focus can provide. As a rule, the outcome of such analyses is likely to make a linguist a more humble scholar: if linguistic research does not restrict itself to the identification of "clear cases" in the grammar (Itkonen 2003, see Willems 2012 for discussion), naturally occurring utterances more often than not show a much greater amount of variation than what one would expect on the basis of one's own linguistic competence. However, this finding is as much in need of explanation as the structure of an individual's knowledge of grammar.

One example may suffice to demonstrate the importance of the issue. For many years now I have been intrigued by phenomena of case variation in German. Time and again I have observed that what you find in large corpora of naturally occurring language is often at variance with what traditional grammars and dictionaries of German claim. Take, for instance, the variation in case marking of the prepositional phrase with the labile verb *aufsetzen* in the meaning 'land on' in present-day German. It is customary (e.g., Ágel 2000: 165) to explain the use of either accusative or dative as follows: In (1) *aufsetzen* is causative and because the sentence is transitive, the prepositional phrase is said to have a 'directional' meaning which triggers the use of accusative (examples from Ágel 2000: 165):

(1) Der Pilot setzte die Maschine sicher **auf die$_{\mathbf{ACC}}$ Piste** auf.
'The pilot landed the aircraft safely on the runway.'

Conversely, it is maintained that *aufsetzen* is a recessive verb in (2) and in accordance with the intransitivity of the sentence the prepositional phrase is said to have a 'locative' meaning and therefore dative marking:

(2) Das Flugzeug setzte **auf dem$_{\mathbf{DAT}}$ Boden** auf.
'The aircraft landed on the ground.'

The proposed explanation can thus be summarized as follows:

	Accusative ('directional')	Dative ('locative')
Intransitive	–	√
Transitive	√	–

Figure 1

But if you browse the Mannheim Deutsches Referenzkorpus (http://www.ids-mannheim.de/cosmas2/), which is a 5,5 billion word corpus of mainly newspaper texts, the empirical findings present a very different picture. The numbers in the chart below record the case variation in an exploratory corpus sample of 361 randomly selected example sentences with *aufsetzen* in the meaning 'land on' in present-day German:

	Accusative	Dative	TOTAL
Intransitive	3	312	315
Transitive	4	42	46
TOTAL	7	354	361

Figure 2

Not only do we see that the intransitive construction with *aufsetzen* in this particular meaning is much more common than the transitive construction, we also find that dative is much more frequent than accusative even in transitive sentences (see, e.g., 3). Conversely, although few in number, accusative is attested in intransitive sentences as well (see, e.g., 4).

(3) Schließlich gelingt es den Piloten wie durch ein Wunder, das Flugzeug sicher **auf der$_{\mathbf{DAT}}$ Landebahn** des Moskauer Militärflughafens Tschkalowo aufzusetzen.
'As if by magic, the pilot finally succeeds in landing the airplane safely on the runway of the Chkalovsky military airport in Moscow.'

(4) Es ist 23.27 Uhr am Dienstag, Ortszeit Rabat, als das Flugzeug **auf die$_{\mathbf{ACC}}$ Landebahn** aufsetzt.
'It is Tuesday 23.27 o'clock, Rabat local time, when the plane lands on the runway.'

These observations squarely contradict the traditional explanation of the case alternation according to which no dative marking is to be expected in the transitive construction and no accusative marking in the intransitive construction. The conclusion must be that a seemingly clear-cut instance of case distribution, which according to traditional accounts that are not based on corpus-data can be explained by means of a straightforward pair of dichotomies, viz. intransitive + dative (locative) vs. transitive + accusative (directional), turns out not to be so once corpus data are taken into consideration. Although the correlation between transitivity and case marking is still statistically significant with regard to our sample of 361 sentences (χ^2

= 8.9121, df = 1, p-value = 0.003, bearing in mind, however, the low number of sentences with accusative; Fisher-Exact test: p-value = 0.006), the bare observations in Figure 2 cast serious doubt on the linguistic validity of the presumed relationship between the morphosyntactic difference between the transitive and intransitive structure, the semantic categories 'directional' and 'locative' and the choice for either accusative or dative.

Findings such as these have potentially important implications for the general debate about the relationship between grammar and usage, and in particular the role of (token) frequency in this discussion. There can be no doubt that it is part and parcel of the German grammatical system that *aufsetzen* strongly correlates with a prepositional phrase in the dative when used intransitively in the conventionalized sense of 'land on'. But at the same time the aforementioned quantitative observations indicate that the use of the "normal" dative in this particular syntactic and semantic environment competes with the use of the accusative which however appears to be much less common, to the extent that even in transitive sentences, instances of dative marking by far outnumber occurrences of accusative marking. There is obviously no point in trying to resolve the issue by simply postulating, on the one hand, that there is a rule of German grammar according to which intransitive *aufsetzen auf* 'land on' governs the dative and takes on a locative meaning, whereas transitive *aufsetzen auf* 'land on' governs the accusative and takes on a directional meaning, while on the other hand relegating all intransitive instances governing the accusative and all transitive instances governing the dative to ultimately extra-grammatical, "deviant" variation in language use. Such a dichotomous explanation, apart from being falsified by the data, ignores the fact that positing such a rule of grammar would be arbitrary from the outset. What is more, even if certain native speakers of German would subscribe to this rule, and hence endorse the ensuing dichotomous explanation because in their judgment it adequately captures the case alternation with intransitive/transitive *aufsetzen* 'land on' in their idiolect, this would still not prove its validity *because it is not supported by the empirical evidence in the corpus data*. But does this entail that the distinction between grammar and usage is spurious or altogether dispensable? Does it entail that an account of usage data based on probabilities and statistically significant preferences can ultimately supplant, rather than supplement, the scholarly identification of grammatical rules?

Newmeyer (2003, 2005, 2006) has forcefully expressed the view that corpus-based frequency studies can add nothing to our understanding of the grammar of any particular individual, although he recognizes that "language users and hence their grammars are sensitive to frequency" (Newmeyer 2003: 697, see also 2006: 705). According to Newmeyer, "variationist work on the grammar/usage interface does not entail as a necessary consequence that properties of grammars and statistical generalizations about the use of constructs provided by grammars are part and parcel of the same cognitive system" (Newmeyer 2006: 706). In contrast, authors such as Guy (2005) maintain that speakers not only are sensitive to frequencies but that any model of linguistic competence has to incorporate the kind of probabilistic information that can be extracted from variationist research (cf. also Clark 2005, Meyer/Tao 2005, Laury/Tsuyoshi 2005). To my mind, with due qualifications one can agree with both these points of view without contradiction.

On the one hand, any realistic grammar of a language has to acknowledge the existence of several kinds of variation, that is, not only variation resulting from the various ways

grammatical rules are instantiated in language use (on which, to my knowledge, all linguists agree) but also variation as an intrinsic feature of grammar itself. The latter kind of variation is what Guy (2005: 561), following Weinreich, Labov & Herzog (1968), calls "inherent variability", which he further distinguishes from "orderly heterogeneity", its counterpart in discourse. Succinctly put: there simply is no perfectly consistent grammatical system apart from the abstract system linguists offer of a natural language.

On the other hand, I concur with Newmeyer (2003, 2005, 2006) that acknowledging speakers' sensitivity to frequencies does not invalidate the distinction between grammar and language use. Without this distinction linguistic accounts are bound to be confusing and self-contradictory. It should not come as a surprise, for example, that in Hymesian ethnographic research of communicative practices in different socio-cultural settings the aforementioned distinction is typically maintained:

> Speakers of a language in particular communities are able to communicate with each other in a manner which is not only correct but also appropriate to the socio-cultural context. This ability involves a shared knowledge of the linguistic code as well as of the socio-cultural rules, norms and values which guide the conduct and interpretation of speech. (Farah 1997: 125).

One can describe but not explain empirically observable language variation without assuming that any one speaker, whenever engaged in the activity of speech, calls upon shared knowledge of a "systematic nature". This knowledge is multifaceted, as Farah (1997) rightly points out in the above quote, but it certainly includes knowledge of language-specific phonemic, prosodic, morphological and syntactic rules, all of which are co-extensive with the knowledge of language-specific meanings ("signifiés" in the sense of Saussure, as opposed to world knowledge and conventionalized senses).

However, it is important to stress that the distinction which supporters as well as detractors draw between (knowledge of) grammar and language use pertains to the domain of *linguistic inquiry* and not to the domain of *language* as the object of inquiry. Viewed from that angle, the claim in modern linguistics that the rigid distinction between the Saussurian *langue* and *parole* has to be overcome is absolutely justified, and corpus-based and variationist research is certainly among the most promising avenues to meet this challenge. But this requires a theoretically consistent view of language as a creative human activity alongside the premises of its scientific inquiry in linguistic research. Coseriu (1973), in his still unsurpassed treatise on the theory of historical linguistics, explains this relationship as follows:

> Por consiguiente, siendo ἐνέργεια en el sentido humboldtiano y aristotélico, el hablar es idealmente anterior a la "lengua" y su objeto (que es la significación) es necesariamente infinito. En este sentido, el lenguaje no se define satisfactoriamente cuando se dice que es "la actividad que *emplea* signos [ya hechos]": hay que definirlo como "actividad *creadora* de signos". Eso, idealmente. Históricamente, en cambio, la "potencia" es anterior al "acto". Hay que integrar, pues, la libertad con la historicidad: en cuanto actividad histórica, el hablar es siempre hablar una "lengua", que es δύναμις histórica; y, en cuanto actividad libre, el hablar no depende enteramente de su potencia, sino que la supera. […] La lengua es una "abstracción" solo técnicamente, para el lingüista que la deduce de la actividad lingüística, y, si puede "abstraerse", es porque existe (como modo de hablar y como saber lingüístico) y porque ya al empezar su estudio tenemos el "conocimiento previo" de su objetividad. Por otra parte, y

> contrariamente a lo que a menudo se piensa, el reconocer la objetividad de la "lengua" y el estudiarla como tal no significa "aislarla" o "separarla" del hablar. El positivismo lingüístico, por su tendencia a "cosificar" las abstracciones, llega, en efecto, a considerar la "lengua" y el "habla" como dos *cosas* distintas y, en lugar de colocar la lengua en el hablar, coloca el "habla" en los individuos y la "lengua" en la sociedad (o peor, en la "masa"), como si los individuos fueran asociales y la sociedad fuera independiente de los individuos y de sus relaciones interindividuales. (Coseriu 1973: 47-50)

In other words, while discourse ("language use") is the source of language ("grammar") in terms of the "conditions of possibility" (Kant) of language, historically discourse is the realization of grammar, i.e. grammatical knowledge. Discourse is emphatically not simply "using" rules of grammar (cf. Coseriu 1985; 2007: 69–75). It therefore seems appropriate to rephrase Newmeyer's famous dictum 'grammar is grammar and usage is usage' in order to overcome its inherent dichotomous perspective: 'grammar is part of discourse, and discourse creates grammar'.

This point of view resolves a seeming paradox. On the one hand, corpus data are much richer than the speech production of a single speaker; on the other hand, what native speakers can produce on the basis of their language competence is richer than anything we will ever find in corpora. This is so because "language as ἐνέργεια is infinite" (Coseriu) and because actual discourse always goes beyond that which generations of speakers have already produced; at the same time the number and variation of structures in texts included in corpora of course exceed the discourse produced by any one individual. Corpora provide data that native speakers may not be aware of (Meyer/Tao 2005: 228); but corpora cannot be used to prove what is *not* possible in a language. Let me briefly touch on the consequences of this position for the discussion about the status of frequency data in linguistic explanations.

In a stimulating discussion about the value of the notion of iconicity in language and linguistics, Haiman (2008) has taken issue with Haspelmath's (2008a, b) argument that the concept of diagrammatic iconicity can be dispensed with in favour of frequency when explaining certain linguistic phenomena. Haspelmath acknowledges that iconicity can play a motivating role in language, e.g., as when the sequence of forms matches the sequence of experience or when forms that belong together tend to occur next to each other (Haspelmath 2008a: 3). However, other structures which seem to be iconic are in fact motivated by frequency-induced ("Zipfian") reduction, according to Haspelmath. For example, no appeal to iconicity is necessary to account for the fact that in most languages the singular often is morphologically shorter than the plural. Haspelmath's alternative explanation is based on the assumption that a language system is both "economic" and "efficient" (Haspelmath 2008a: 5) and that "frequency of use implies short coding because frequent items are more predictable" (Haspelmath 2008b: 59). But note that one crucial question remains disturbingly unanswered: How does it come that the most frequent forms, which are the most predictable ones, are also the shortest ones? Obviously, the answer cannot be that this is so because they are the most frequent and predictable ones, that would be circular.[1] Haiman (2008: 36) is right to point out that frequency is no "cognitive" explanation whatsoever in linguistics. But no solution to the

[1] To be sure, the most frequent forms are the most predictable ones only from a dictionary point of view. From the point of view of discourse (i.e. the "text" with its typical properties of formal cohesion, semantic coherence, isotopy, information structure etc.), a form which is very likely to occur in the next sentence and which in this sense is highly predictable, may be infrequent in the language.

problem can be obtained without addressing a potential source of confusion: frequency – that is, token frequency – does not pertain to discourse, it is a feature of the product of discourse, i.e. the body of texts produced in a particular language. Therefore, frequency in itself cannot possibly account for why certain forms are shorter than other ones, although it is an important finding of quantitative research of texts that a correlation between frequency, predictability and length of forms can be identified and measured. Hence not the frequency but only the pervasive intention to produce shorter forms for particular functions (which have to be "meaningful") in discourse can explain any frequency-induced reduction of linguistic forms that is observable in texts. Once again, I quote extensively from Coseriu (1973), because what he writes about the paradox of language change equally applies to the apparent "causality" of frequency in order to account for – or, better still, dismiss – iconicity in language:

> En el fondo, la perplejidad frente al cambio lingüístico y la tendencia a considerarlo como fenómeno espurio, provocado por "factores externos", se deben al hecho de partir de la lengua abstracta – y, por lo tanto, estática –, separada del hablar y considerada como cosa hecha, como *ergon*, sin siquiera preguntarse qué son y cómo existen realmente las lenguas y qué significa propriamente un "cambio" en una lengua. De acquí también el planteamiento del problema del cambio en términos causales, puesto que los cambios en las "cosas" desligados de la intencionalidad de todo sujeto se atribuyen, precisamente, a "causas". Pero la lengua no pertenece al *orden causal* sino al *orden final*, a los hechos que se determinan por su función. (Coseriu 1973: 29–30)

Coseriu's Humboldtian conception of language as "actividad *creadora* de signos" rather than an "actividad que *emplea* signos [ya hechos]" also resolves a vexing theoretical problem that any concept of "emergent grammar" such as that advocated by Paul Hopper is faced with. On the one hand, Hopper correctly points out that the grammar of a language is "constantly being restructured and resemanticized during actual use" (Hopper 1998: 159); emergence in this sense designates "a continual movement toward structure" (157). On the other hand, Hopper takes his rejection of the dichotomy between grammar and usage to be equivalent to the view that linguistic structure is "epiphenomenal", that is, "an effect rather than a cause" (157), and he defines emergent regularities as "the sediment of frequency" (161; cf. also 158). But this brings us back to the much maligned dichotomy, albeit on a more abstract level, rather than overcoming it. An epiphenomenal effect of language use that results in a sediment of forms and structures is precisely the kind of *thing* ("cosa") that, severed from its sources (i.e., various discourse practices), may be construed as the 'static' prerequisite of 'dynamic' language use.[2] In this sense, theoretically separating language use from grammar amounts to objectifying language in order to make it amenable to causal explanations. However, whatever the merits are of establishing causal relationships with respect to classifying various quantitative features of texts and even predicting the likelihood with which a particular structure occurs under specific conditions (see, e.g., Bresnan & Ford 2010), causal explanations do not, and cannot, fully explain the intentional, and fundamentally historical, activity human beings are engaged in when they create language.

[2] Note that Newmeyer's position is diametrically opposed to what Hopper claims. Newmeyer (2006: 402) writes that probabilities are not part of knowledge of grammar but "totally epiphenomenal". In the approach that I advocate, Hopper's and Newmeyer's views are equally problematic on this point.

To round off this discussion note, let us return to our initial example of case alternation in the context of prepositional constructions with verbs such as *aufsetzen* in German. Postulating simple dichotomous rule-sentences which are based on apparently simplifying interpretations of a few artificially constructed examples along the lines of traditional grammars and dictionaries is certainly not going to help us in finding an explanation of the observed variation between accusative and dative. I dare say that no native speaker of German has ever come up with an entirely satisfactory explanation of the case alternation, in spite of being fully competent in the language and not feeling the urge to account for something that most likely does not constitute any hindrance to communication most of the time. However, this does not invalidate scholarly endeavours that set out to accommodate the observed variation in the grammar. Likewise, corpus-based findings that clearly contradict traditional accounts of grammatical variation do not by themselves entail that the distinction between grammar and usage is dispensable, nor that an account of usage data based on probabilities and statistically significant preferences can simply supplant the formulation of grammatical rules. Quite on the contrary, such findings are an invitation to reconsider already existing explanations and explore alternative ones that better fit the facts. With respect to the observed case alternation of prepositional constructions with verbs such as *aufsetzen* in German, for instance, it is imperative to take into consideration various factors that have received no or only scant attention in the past, including paradigmatic and syntagmatic contrasts, correlations between constructions of varying complexity and case preferences, defeasible default case marking in relation to conventionalized senses at the level of "normal language use" (Coseriu 1973: 53–57), etc. (cf. Willems 2011 and Willems, Rys, De Cuypere Forthc. for more elaborate accounts). To the extent that multifactorial quantitative analyses are able to capture subtle interactions between a diverse array of factors, one may be confident that their empirical findings will greatly benefit the ensuing grammatical analyses without falling prey to a naive positivism which confuses descriptive detail with a comprehensive understanding of language in discourse.

References

Ágel, Vilmos (2000): *Valenztheorie*. Tübingen: Gunter Narr Verlag.

Bresnan, Joan/Ford, Marilyn (2010): "Predicting syntax: Processing Dative Constructions in American and Australian varieties of English", in: *Language* 86:1, 186–213.

Clark, Brady (2005): "On stochastic grammar", in: *Language* 81:1, 207–217.

Coseriu, Eugenio (1973): *Sincronía, diacronía e historia. El problema del cambio lingüístico*. Madrid: Gredos (1st ed. 1958).

Coseriu, Eugenio (1985): "Linguistic competence: What is it really?", in: *The Modern Language Review* 80:4, xxv-xxxv.

Coseriu, Eugenio (2007): *Sprachkompetenz. Grundzüge der Theorie des Sprechens.* Tübingen: Gunter Narr (1st ed. 1988).

Farah, Iffat (1997): "Ethnography of communication", in: N. H. Hornberger/ D. Corson (eds.): *Encyclopedia of language and education*. Vol. 8: *Research methods in language and education*. Dordrecht: Kluwer, 125–133.

Guy, Gregory R. (2005): Letter to *Language* 81:3, 561–563.

Haiman John (2008): "In defence of iconicity", in: *Cognitive Linguistics* 19:1, 35–48.

Haspelmath, Martin (2008a): "Frequency vs. iconicity in explaining grammatical asymmetries", in: *Cognitive Linguistics* 19:1, 1–33.

— (2008b): "Reply to Haiman and Croft", in: *Cognitive Linguistics* 19:1, 59–66.

Hopper, Paul J. (1998): "Emergent grammar", in: M. Tomasello (ed.): *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Lawrence Erlbaum, 155–175.

Itkonen, Esa (2003): *What is Language? A Study in the Philosophy of Linguistics*. Turku: Turun yliopisto (*Publications in General Linguistics*, vol. 8).

Laury, Ritva /Tsuyoshi, Ono (2005): "Data is data and model is model: you don't discard the data that doesn't fit your model!", in: *Language* 81:1, 218–225.

Meyer, Charles F./ Tao, Hongyin (2005): "Responce to Newmeyer's 'Grammar is grammar and usage is usage'", in: *Language* 81:1, 227–228.

Newmeyer, Frederick J. (2003): "Grammar is grammar and usage is usage", in: *Language* 79:4, 682–707.

— (2005): "A reply to the critiques of 'Grammar is grammar and usage is usage'", in: *Language* 81:1, 229–236.

— (2006): "Grammar and usage: A response to Gregory R. Guy", in: *Language* 82:4, 705-706.

Weinreich, Uriel/Labov, William/Herzog, Marvin Herzog (1968): "Empirical foundations for a theory of language change", in: W. Lehmann/Y. Malkiel (eds.): *Directions for historical linguistics*. Austin: University of Texas Press, 95–188.

Willems, Klaas (2011): "The semantics of variable case marking (Accusative/Dative) after two-way prepositions in German locative constructions: towards a constructionist approach", in: *Indogermanische Forschungen* 116, 324–266.

— (2012): "Intuition, introspection and observation in linguistic inquiry", in: *Language Sciences* 34:6, 665–681.

Willems, Klaas/Rys, Jonah/De Cuypere, Ludovic Forthc.: "Case alternation in argument structure constructions with prepositional verbs. A case study in corpus-based constructional analysis", in: H. Boas/A. Ziem (eds.). *Constructional approaches to argument structure in German* [working title]. Berlin & New York: Mouton de Gruyter.